PART II – PROBABILISTIC FUZZY MODELS

Fuzzy systems

Many successful applications and solutions
- Data analysis and modelling
- Image processing
- Control systems
- Household appliances
- Consumer electronics
- Multi-agent system design
- Decision making under uncertainty
- Information retrieval
Deterministic output, probabilistic uncertainty

Density estimation and fuzzy systems

- After defuzzification, fuzzy systems implement a nonlinear mapping from inputs to outputs.
- Output is assumed to be deterministic: regular fuzzy systems cannot estimate densities.
- Many problems require estimation of densities: extend fuzzy systems to estimate density.
- In doing so, linguistically interpretable models can also be developed for probability density estimation → probabilistic fuzzy systems.
Applications of density estimation

- Computing error bounds on models
- Value-at-risk estimation for financial risks
- Modeling stochastic outcomes in healthcare
- Linguistic descriptions of probability distributions
- Linguistic descriptions for stochastic time series
- Etc.

Probabilistic vs. fuzzy systems

**Probabilistic systems**
- Consider uncertainty as randomness
- Emphasis on statistical properties of data
- Axiomatic grounding
- Assumptions often taken as a priori

**Fuzzy systems**
- Emphasis on linguistic uncertainty
- Statistical properties of data often ignored
- Function approximation properties (deterministic)
- Focus on linguistic grounding
Hammer principle

If your only tool is a hammer, you start seeing the whole world as a nail
(H.-J. Zimmermann)

Combining probability and fuzziness

- There have been many approaches to combine probability and fuzziness
  - Fuzzy random variables, random fuzzy variables, fuzzy probabilistic systems, probabilistic fuzzy systems, certainty factor methods, probabilistic rule weighting, etc.
- Many contributions, a.o. from
Probability of a fuzzy statement

What is the probability that a randomly selected Indian person is very tall?

- Probability is unlikely (low, small, etc.)
- Probability is about 0.4
- Probability is 0.422

Preliminaries

Probability of a fuzzy event:

- Crisp
  \[ \Pr(A) = \int_{x \in A} p(x) \, dx = \int_X \mathcal{X}_A(x) p(x) \, dx \]
- Fuzzy (Zadeh, 1968)
  \[ \Pr(A) = \int_X \mu_A(x) p(x) \, dx \]

Conditional probability:

\[ \Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\int_X \mu_A(x) \mu_B(x) p(x) \, dx}{\int_X \mu_B(x) p(x) \, dx} \]
Preliminaries (2)

Probability of a fuzzy event:

- Crisp
  \[ \Pr(A) = \int_{x \in A} p(x) \, dx = \int_X \chi_A(x) p(x) \, dx \]

- Fuzzy (Zadeh, 1968)
  \[ \Pr(A) = \int_X \mu_A(x) p(x) \, dx \]

Conditional probability:
\[
\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\int_X [\mu_A(x) \wedge \mu_B(x)] p(x) \, dx}{\int_X \mu_B(x) p(x) \, dx}
\]

Satisfies \( \Pr(A \mid A) = 1 \)

An example

Answering the question:

“What is the probability that a randomly selected Indian woman is tall?”

Calculating the probability using a given pdf of length.
Simple estimation

- Let \( x_1, \ldots, x_n \) be a random sample on a domain \( X \)
- The probability of a crisp event \( A \) can be estimated by
  \[
  \frac{1}{n} \sum_{k} \chi_A(x_k)
  \]
- The probability of a fuzzy event \( A_i \) can be estimated by
  \[
  \frac{1}{n} \sum_{k} \mu_{A_i}(x_k)
  \]
  assuming that \( X \) is well-formed, i.e.
  \[
  \forall x_k \sum_{i} \mu_{A_i}(x_k) = 1
  \]

Deterministic and probabilistic rules

Rule \( R_q \): If \( x \) is \( A_q \) then \( y \) is \( C_q \)

Model parameters: \( A_q, C_q \)

If current returns are large, then future returns will be large

If current returns are large, then future returns will be large with probability \( \rho \).
Future returns will be small with probability \( 1 - \rho \).
Probabilistic fuzzy rules

Rule $R_q$: If $x$ is $A_q$ then
- $y$ is $C_{1q}$ with $\Pr(C_{1q}|A_q)$ and...and
- $y$ is $C_{j,q}$ with $\Pr(C_{j,q}|A_q)$ and...and
- $y$ is $C_{k,q}$ with $\Pr(C_{k,q}|A_q)$

Model parameters:
- $A_q$, $C_{j,q}$, $\Pr(C_{j,q}|A_q)$

For practical purposes, one can assume that
$$\forall j, q, q' \quad C_{j,q} = C_{j,q'}$$

Probabilistic fuzzy systems

- Consists of a set of rules whose antecedents are fuzzy conditions and whose consequents are probability distributions

Rule $R_j$:
- antecedent
- consequent
- fuzzy set

where,
$$\Pr(\omega_c|A_j) \geq 0, \text{ for } j = 1, \ldots, J \text{ and } c = 1, \ldots, C$$
and
$$\sum_{c=1}^{C} \Pr(\omega_c|A_j) = 1, \text{ for } j = 1, \ldots, J.$$ (3)
Deterministic vs. probabilistic FS

If $x$ is $A^4$ then $y$ is $B^3$ with probability $p(B^3 | A^4)$,
and $y$ is $B^2$ with probability $p(B^2 | A^4)$,
and $y$ is $B^3$ with probability $p(B^3 | A^4)$.

Probabilistic fuzzy system

Rule $R_q$: If $x$ is $A_q$ then $f(y) = f(y | A_q)$

Additive reasoning:

$$f(y | x) = \frac{\sum_{q=1}^{Q} \mu_{A_q}(x) f(y | A_q)}{\sum_{q=1}^{Q} \beta_q(x) f(y | A_q)} = \sum_{q=1}^{Q} \beta_q(x) \mathbb{E}(y | A_q)$$

$$y = \mathbb{E}(y | x) = \sum_{q=1}^{Q} \beta_q(x) \mathbb{E}(y | A_q)$$
Probability distribution characterization

- In general, different characterizations can be used for the conditional probability density in the rule consequents.
- This characterization could be an approximation with a histogram or an explicit model for density, e.g., a linear model, a GARCH model, etc.
- In PFS, we can select a fuzzy histogram characterization.

Histograms

- Let \( x_1, \ldots, x_n \) be a random sample from a univariate distribution with pdf \( f(x) \).
- Let the characteristic functions \( \chi_i(x) \) (defining crisp bins/intervals \( A_i \)) constitute a crisp partitioning:
  \[
  \forall x \in \mathbb{R} : \sum_i \chi_i(x) = 1 \quad \forall i \neq j : \chi_i(x) \chi_j(x) = 0
  \]
- A histogram estimates \( f(x) \) as follows:
  \[
  \hat{f}(x) = \sum_i \frac{\chi_i(x) p_i}{|A_i|} \approx \sum_i \frac{\chi_i(x) \left( \frac{1}{n} \sum_k \chi_i(x_k) \right)}{\int \chi_i(x) dx}
  \]
Fuzzy histograms

- Let \( x_1, \ldots, x_n \) be a random sample of size \( n \) from a (univariate) distribution with pdf \( f(x) \)
- Let the membership functions \( \mu_i(x) \) (defining fuzzy bins \( A_i \)) constitute a fuzzy partitioning:

\[
\forall x \in \mathbb{R} : \sum \mu_i(x) = 1
\]

- A fuzzy histogram estimates \( f(x) \) as follows:

\[
f(x) = \sum_i \frac{\mu_i(x)p_i}{|A_i|} \approx \sum_i \frac{\mu_i(x)\left(\frac{1}{n} \sum_k \mu_i(x_k)\right)}{\int \mu_i(x)dx}
\]

Crisp vs. fuzzy histogram

- Membership values
- pdf normal distribution
- Crisp histogram
- Fuzzy histogram
Two interpretations

- Original interpretation:
  \[ \hat{f}(x) = \sum_{i} \frac{\mu_i(x) p_i}{\int \mu_i(x) dx} = \sum_{i} \mu_i(x) \int \frac{p_i(x)}{\mu_i(x)} dx \]
  concerns an interpolation between local densities

- Another interpretation:
  \[ f(x) = \sum_{i} \frac{\mu_i(x) p_i}{\int \mu_i(x) dx} = \sum_{i} \mu_i(x) \int \frac{p_i(x)}{\mu_i(x)} dx \]
  concerns a sum of pdfs, weighted with \( p_i \)

Relation to kernel density estimation

- Fuzzy histograms are similar to Parzen Windows for density estimation
- There is also a close relation with kernel density estimation

Kernel density estimator:
\[ \hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K_h(x-x_j) \]

Fuzzy histogram:
\[ f(x) = \frac{1}{n} \sum_{j=1}^{n} K_h(x,x-x_j), \quad \text{with} \]
\[ K_h(x,x-x_j) = \frac{1}{h} \sum_{i \in Z} \mu(\frac{x}{h} - i) \mu(\frac{x_j}{h} - i). \]

So, a FH is a KDE that uses different kernels for different values of \( x \)
Fuzzy histogram model

- Replace the true pdf $f(y|A_q)$ by a fuzzy histogram $\hat{f}(y|A_q)$

$$\hat{R}_q : \text{if } x \text{ is } A_q \text{ then } f(y) \text{ is } \hat{f}(y|A_q),$$

where $\hat{f}(y|A_q)$ is defined as

$$\hat{f}(y|A_q) = \sum_{j=1}^{J} \frac{Pr(C_j|A_q)\mu_{C_j}(y)}{\int_{-\infty}^{\infty} \mu_{C_j}(y)dy}.$$

- The expected output and variance are

$$\hat{\mu}_{y|x} = \hat{E}(y|x) = \sum_{q=1}^{Q} \sum_{j=1}^{J} \beta_q(x) \hat{Pr}(C_j|A_q) z_j,$$

$$\hat{\sigma}_{y|x}^2 = \hat{E}(y^2|x) - (\hat{E}(y|x))^2 = \sum_{q=1}^{Q} \sum_{j=1}^{J} \beta_q(x) \hat{Pr}(C_j|A_q) \zeta_j - \hat{\mu}_{y|x}^2,$$

where $\beta_q(x)$ is the normalized degree of fulfillment, $z_j$ is the centroid of the $j$th output fuzzy set and $\zeta_j = \frac{\int_{-\infty}^{\infty} y^2 u_{C_j}(y)dy}{\int_{-\infty}^{\infty} u_{C_j}(y)dy}.$
Probabilistic Mamdani systems

Rule $R_q$: If $x$ is $A_q$ then

$\hat{y} = C_1$ with $\Pr(C_1|A_q)$ and...and

$\hat{y} = C_j$ with $\Pr(C_j|A_q)$ and...and

$\hat{y} = C_J$ with $\Pr(C_J|A_q)$

Reasoning:

$y = E(y|x) = \sum_{q=1}^{Q} \sum_{j=1}^{J} \beta_q(x) \Pr(C_j|A_q) z_j$

Centroid of fuzzy consequent set $C_j$

$z_j = \frac{\int_{y} y \mu_{C_j}(y) dy}{\int_{y} \mu_{C_j}(y) dy}$

Constrained consequent partition

Rules specify a probability distribution over a collection of fuzzy sets that partition the output domain

$\sum_{j=1}^{J} \mu_{C_j}(y) = 1, \forall y \in Y$

Additive reasoning with multiplicative aggregation of rule antecedents

$f(y|x) = \sum_{j=1}^{J} \sum_{q=1}^{Q} \beta_q(x) \Pr(C_j|A_q) \frac{\mu_{C_j}(y)}{\int_{-\infty}^{\infty} \mu_{C_j}(y) dy}$
Probabilistic fuzzy output model

Rule \( \hat{R}_q \): If \( x \) is \( A_q \) then \( y \) is \( C_1 \) with \( \hat{P}(C_1 | A_q) \) and

\[
\vdots
\]

\( y \) is \( C_J \) with \( \hat{P}(C_J | A_q) \).

Example pdf estimated from PFS
Probabilistic TS systems

Zero-order probabilistic Takagi-Sugeno

Rule $R_q$ : If $x$ is $A_q$ then

\[
y = y_1 \text{ with } \Pr(y_1|A_q) \quad \text{and}
y = y_2 \text{ with } \Pr(y_2|A_q) \quad \text{and} \ldots \text{and}
y = y_j \text{ with } \Pr(y_j|A_q)
\]

In essence, consequents are crisp sets centered around $y_j$

Reasoning: $y = E(y | x) = \sum_{q=1}^{Q} \sum_{j=1}^{J} \beta_q(x) y_j \Pr(y_j|A_q)$

Relation to deterministic FSs

- Zero-order Takagi-Sugeno system

Takagi-Sugeno reasoning $y^* = \sum_{q=1}^{Q} \beta_q(x) c_q$

c.f. $y = E(y | x) = \sum_{q=1}^{Q} \sum_{j=1}^{J} \beta_q(x) \Pr(y_j|A_q) z_j$

Select $c_q = \sum_{j=1}^{J} \Pr(y_j|A_q) z_j$
Relation to CF fuzzy classifiers

- Fuzzy classifiers with certainty factor reasoning:
  
  \( R_q : \text{If } x \text{ is } A_q \text{ then class is } c \text{ with } CF_q \)
  
  By selecting certainty factors appropriately, good classification results can be obtained

- It has been studied that:
  
  - CF approach uses *maximum of terms* while PFS approach takes the *sum of terms* (additive reasoning)
  
  - in cases at most two rules overlap, the decision rules of CF and PFS approaches yield the same result

Van den Berg et al. (2002)

Probabilistic fuzzy systems

- Essentially a fuzzy system that estimates a probability density function, i.e. the fuzzy system approximates a p.d.f.
- Usually p.d.f. is conditional on the input
- Linguistic information is coded in fuzzy rules
- Related to density estimation techniques such as Parzen windows and kernel based density estimation
- Combine linguistic uncertainty with probabilistic uncertainty
- Different types of fuzzy systems can be extended to the PFS equivalent (e.g. Mamdani fuzzy systems, Takagi-Sugeno fuzzy systems)
PFS design

- Identifying mental world vs. observed world (van den Eijkel 1999)
- Mental world: linguistic descriptions, fuzzy conceptualization, experts’ knowledge
- Observed world: data measurements, probability density functions, optimal consequent parameters
- Optimal design given a mental world: application of conditional probability measures for fuzzy events
- Optimal design given an observed world: nonlinear optimization techniques

Parameter determination

<table>
<thead>
<tr>
<th>Sequential Method</th>
<th>Maximum Likelihood Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1 MF parameters</td>
<td>Part 1 – Initialization</td>
</tr>
<tr>
<td>Part 2 Probability parameters</td>
<td>Sequential Method</td>
</tr>
<tr>
<td>Part 2 - Optimization</td>
<td>Part 2 - Optimization</td>
</tr>
</tbody>
</table>

MF – Membership function
**Sequential method**

- **Part 1**: Finding the membership function (MF) parameters

- $A_i$: a fuzzy set defined by a membership function $\mu_{A_i}(x)$

- E.g. – Gaussian MF parameters:
  $$\phi(x) = \exp\left(-\sum_{n=1}^{N} \frac{(x_n - \bar{x}_n)^2}{\sigma_n^2}\right)$$
  - $\bar{x}$ – center of MF
  - $\sigma$ – width of MF

**MF determination**

- For the first part of the sequential method, well-known techniques from fuzzy modeling can be applied
  - Fuzzy clustering in input-output product space
  - Fuzzy clustering in input and output space
  - Expert-driven design
  - Similarity-based rule-base simplification
  - Feature selection
  - Heuristics approaches
  - Etc.
Sequential method

- Part 2: Finding the probability parameters - $Pr(\omega_c|A_j)$

\[ y = \omega_1 \text{ with probability } Pr(\omega_1|A_j) \text{ and } \]
\[ y = \omega_2 \text{ with probability } Pr(\omega_2|A_j) \text{ and } \]
\[ ... \]
\[ y = \omega_C \text{ with probability } Pr(\omega_C|A_j) \]

- Set the parameters $Pr(\omega_i/A_j)$ equal to estimates of the conditional probabilities - conditional probability estimation

\[
Pr(\omega_c|A_j) = \frac{\sum_{i=1}^{I} \omega_{A_i}(x_i) \chi_{\omega_c}(y_i)}{\sum_{i=1}^{I} \beta_{A_j}(x_i)}
\]

Estimation of probability parameters

- Conditional probabilities $Pr(C_j | A_q)$ can be assessed directly by using the definition of the probability of joint events:

\[
Pr(C_j | A_q) = \frac{Pr(A_q \cap C_j)}{Pr(A_q)} = \frac{\sum_{(x_k,y_k)} A_q(x_k) C_j(y_k)}{\sum_{x_k} A_q(x_k)}
\]

- This method does not provide maximum likelihood estimates of the probability parameters.
Bias in parameter estimation

Underlying data generation process

\[ \mu_{A_1}(x) = 1 - x \]
\[ \mu_{A_2}(x) = x \]
\[ \mu_{C_1}(y) = 1 - y \]
\[ \mu_{C_2}(y) = y \]

\[ \Pr(y, x) = 2xy - x - y + 1 \]

Optimal parameters:

\[ \Pr(C_1 | A_1) = 1 \quad \Pr(C_1 | A_2) = 0 \]
\[ \Pr(C_2 | A_1) = 0 \quad \Pr(C_2 | A_2) = 1 \]
Maximum likelihood method

- Part 2: Optimization of $v_j$, $\sigma_j$, and $Pr(\omega_c|A_j)$

Likelihood of a data set

$$L = \prod_{i=1}^{t} \hat{P}(y_i|x_i).$$

Minimization of the negative log-likelihood

$$E = -\sum_{i=1}^{t} \ln \hat{P}(y_i|x_i).$$

Optimize parameters $v_j$, $\sigma_j$, and $Pr(\omega_c|A_j)$ that minimize the error function

Constrained optimization problem
(probability parameters $Pr(\omega_c|A_j)$ must satisfy summation conditions)

Constrained optimization problem
Unconstrained optimization problem

$$Pr(\omega_c|A_j) = \frac{e^{u_{jc}}}{\sum_{c=1}^{C} e^{u_{jc}}}.$$

- Unconstrained minimization of $v_j$, $\sigma_j$, and $u_{jc}$
- Gradient descent optimization algorithm is used to minimize the objective function – i.e. the available classification examples are processed one by one and updates are performed after each sample
Experimental comparison (1)

- Use Gaussian membership functions
  \[ \mu_{q_l}(x) = \exp\left( -\sum_{i=1}^{d} \frac{(x_i - c_{q_l})^2}{\sigma_{q_l}^2} \right) \]
- The centers \( c_{q_l} \) are determined using fuzzy c-means clustering
- The widths \( \sigma_{q_l} \) are set equal to \( \sigma_{q_l} = \min_{j \neq j'} |c_q - c_{q'}| \)

Experimental comparison (2)

- Misclassification rates
<table>
<thead>
<tr>
<th></th>
<th>Wisconsin breast cancer</th>
<th>Wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential method</td>
<td>0.261 (0.036)</td>
<td>0.034 (0.048)</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>0.029 (0.021)</td>
<td>0.023 (0.041)</td>
</tr>
</tbody>
</table>

- Calculated using ten-fold cross-validation
- Standard deviations reported within parentheses
### Concluding remarks

- Probabilistic fuzzy systems combine linguistic uncertainty and probabilistic uncertainty

- Very useful in applications where a probabilistic model (pdf estimation) has to be conditioned (or constrained) by linguistic information

- Good parameter estimation methods exist and the added value of these models has been demonstrated in various applications

### Future research directions

- New estimation methods for the model parameters
  - Joint estimation
  - Information-theory based techniques
  - Better optimization methods

- Interaction linguistic knowledge and data-driven estimation

- Optimizing model complexity, model simplification

- Interpretability of probabilistic fuzzy models

- Linguistic descriptions of probability density functions

- Equivalence to other systems: e.g. fuzzy Markov models

- Density estimation using more complex models as rule consequents: e.g. fuzzy GARCH models

- New applications
Fuzzy Markov models

- Consider a PFS in which the output space is the time-shifted version of the input space: let the states be defined in the same way both for inputs and outputs

- Then, it can be shown that PFS is a fuzzy Markov model in which the probability parameters are the transition probabilities of the Markov model

Fuzzy GARCH models

- Define $L$ as the number of IF-THEN rules, with the $l$-th rule:
  
  Rule ($1$): IF $x_{1,t}$ is $F_{11}$ and ... and $x_{n,t}$ is $F_{1n}$, THEN

  $$h_t^l = \alpha_0^l + \sum_{i=1}^{q} \alpha_i^l y_{t-i}^2 + \sum_{j=1}^{p} \beta_j^l h_{t-j}$$

  $$y_t^l \mid x_t, h_t^l \sim N(\mu^l, h_t^l)$$

  where both the unobserved conditional variance $h$ and $y$ are explained by the rules.

- The output of the system is:

  $$y_t = \sum_{l=1}^{L} g_{l,t} y_t^l = \sum_{l=1}^{L} g_{l,t} \sqrt{h_t^l}, \Rightarrow y_t \sim \sum_{l=1}^{L} g_{l,t} \text{NID}(\mu^l, h_t^l)$$
Conditional distributions of simulated data

- Simulated data from previous definition, with $\mu = 0$; leading to normal conditional densities.
- Simulated data from proposed definition, with $\mu = (\mu_1, \mu_2) = (-6, 6)$; leading to bimodal and skewed conditional densities.

Selected bibliography

Selected bibliography